BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37414
doi: 10.3897/biss.3.37414

OPEN ACCESS

Conference Abstract

Long-Term Reusability of Biodiversity and Collection Data using a National Federated Data Infrastructure

Peter Grobe[‡], Maren Gleisberg[§], Birgit Klasen[‡], Juan Carlos Monje[|], Anke Penzlin[¶], Tanja Weibulat [#], Dagmar Triebel[¤]

‡ Zoological Research Museum Alexander Koenig, Bonn, Germany
§ Botanical Garden and Botanical Museum Berlin Dahlem, Berlin, Germany
| State Museum of Natural History Stuttgart, Stuttgart, Germany
¶ Senckenberg Gesellschaft für Naturforschung, Frankfurt, Germany
# German Federation for Biological Data e.V., Bremen, Germany
¤ SNSB IT Center and Botanische Staatssammlung München, Munich, Germany

Corresponding author: Peter Grobe (p.grobe@leibniz-zfmk.de)

Received: 17 Jun 2019 | Published: 26 Jun 2019

Citation: Grobe P, Gleisberg M, Klasen B, Monje J, Penzlin A, Weibulat T, Triebel D (2019) Long-Term Reusability of Biodiversity and Collection Data using a National Federated Data Infrastructure. Biodiversity Information Science and Standards 3: e37414. https://doi.org/10.3897/biss.3.37414

Abstract

GFBio “German Federation for Biological Data” is a data infrastructure and network set up by several research institutions in Germany. It fosters archiving and long-term reusability of research data and provides open and free access via a joint web portal at www.gfbio.org. As part of the working procedures data are semantically enriched and provided via a visualization and analysis tool. The main aim of the infrastructure is to make research data from the biological domain reusable and accessible on the long run following FAIR principles.

In order to achieve this, several workflows and best practices have been established. The archiving of biodiversity and collection research data follows the reference model (ISO 14721) for an **Open Archival Information System (OAIS)**. The challenges for making data reusable is on the one hand the heterogeneity of this data, on the other hand the often implicit but differing semantics making data integration a hard and difficult process. The

© Grobe P et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

use of data management plans is one approach we run to face and solve the challenges. Data management plans contain recipes about the research data, the tools used to acquire data, the content- and exchange formats, the metadata required to describe the data, and finally the costs and resources needed by data providers to deliver structured “Submission Information Packages” (SIPs) in the sense of OAIS.

The archiving of a data package as “Archival Information Package” (AIP) is not sufficient to make it reusable in the future. Changes in the semantic meaning over time (content obsolescence), changes in the formats (format obsolescence), and changes in the technology of storage media (hardware obsolescence) are the major factors to be considered here. According to the FAIR principles and to our understanding data is best preserved if it is visible and available for use. The **biodiversity and collection data centers** involved in GFBio therefore have **a curation layer** (cf. management aka OAIS) in the archiving pipeline assembling their in-house management systems for sample and observation data and their asset management systems for all kinds of multimedia. This layer allows a continuous quality control and review of the incoming information packages. Thus, data providers can continuously maintain their data if wished for. The data are stored as AIPs sensu OAIS at the specialized data centers and are accessed by GFBio's core system. Dissemination Information Packages (DIPs) can be generated continuously at every time from the data and disseminated using content standards for data and metadata, like EML, ABCD and MIxS. Data are available via the GFBio website and in parallel using other web portals from the biological domain, e.g. INSDC and GBIF.

The GFBio data centers now strive for **certification** of their archiving processes using the Core Trust Seal and for the certification of the FAIRness of single data records.

Established data flows and documentation on **best practices** are available under: www.gfbio.org/data-centers.

Presenting author

Peter Grobe

Presented at

Biodiversity_Next 2019

Acknowledgements

We thank the GFBio consortium.

Funding program

German Research Fundation (DFG): Information Infrastructures for Research Data

Grant title

German Federation for Biological Data (GFBio) Phase III